

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

COPY



A1

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12N 15/31, C07K 14/315, 16/12, C12Q 1/68	A2	(11) International Publication Number: WO 98/18931	(43) International Publication Date: 7 May 1998 (07.05.98)
---	----	---	---

(21) International Application Number: PCT/US97/19588

(22) International Filing Date: 30 October 1997 (30.10.97)

(30) Priority Data:
60/029,960 31 October 1996 (31.10.96) US

(71) Applicant (for all designated States except US): HUMAN
GENOME SCIENCES, INC. [US/US]; 9410 Key West
Avenue, Rockville, MD 20850 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): KUNSCH, Charles, A.
[US/US]; 2398B Dunwoody Crossing, Atlanta, GA 30338
(US). CHOI, Gil, H. [KR/US]; 11429 Potomac Oaks Drive,
Rockville, MD 20850 (US). DILLON, Patrick, J. [US/US];
1055 Snipe Court, Carlsbad, CA 92009 (US). ROSEN,
Craig, A. [US/US]; 22400 Rolling Hill Road, Laytonsville,
MD 20882 (US). BARASH, Steven, C. [US/US]; 582 Col-
lege Parkway #303, Rockville, MD 20850 (US). FAN-
NON, Michael [US/US]; 13501 Rippling Brook Drive, Sil-
ver Spring, MD 20850 (US). DOUGHERTY, Brian, A.
[US/US]; 708 Meadow Field Court, Mount Airy, MD 21771
(US).

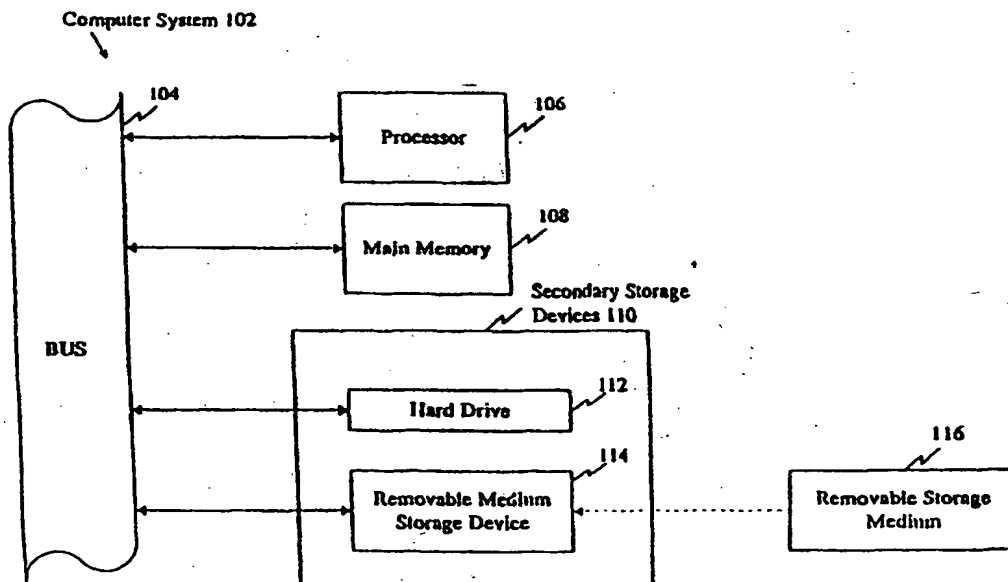
(74) Agents: BROOKES, A., Anders et al.; Human Genome
Sciences, Inc., 9410 Key West Avenue, Rockville, MD
20850 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR,
BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE,
GH, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK,
LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO,
NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR,
TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH,
KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ,
BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE,
CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL,
PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN,
ML, MR, NE, SN, TD, TG).

Published

Without international search report and to be republished
upon receipt of that report.

(54) Title: *STREPTOCOCCUS PNEUMONIAE* POLYNUCLEOTIDES AND SEQUENCES



(57) Abstract

The present invention provides polynucleotide sequences of the genome of *Streptococcus pneumoniae*, polypeptide sequences encoded by the polynucleotide sequences, corresponding polynucleotides and polypeptides, vectors and hosts comprising the polynucleotides, and assays and other uses thereof. The present invention further provides polynucleotide and polypeptide sequence information stored on computer readable media, and computer-based systems and methods which facilitate its use.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EK	Estonia						

presence of *Streptococcus pneumoniae* in a sample, hereinafter referred to as diagnostic fragments or DFs.

Each of the ORFs in fragments of the *Streptococcus pneumoniae* genome disclosed in Tables 1-3, and the EMFs found 5' to the ORFs, can be used in numerous ways as polynucleotide reagents. For instance, the sequences can be used as diagnostic probes or amplification primers for detecting or determining the presence of a specific microbe in a sample, to selectively control gene expression in a host and in the production of polypeptides, such as polypeptides encoded by ORFs of the present invention, particular those polypeptides that have a pharmacological activity.

The present invention further includes recombinant constructs comprising one or more fragments of the *Streptococcus pneumoniae* genome of the present invention. The recombinant constructs of the present invention comprise vectors, such as a plasmid or viral vector, into which a fragment of the *Streptococcus pneumoniae* has been inserted.

The present invention further provides host cells containing any of the isolated fragments of the *Streptococcus pneumoniae* genome of the present invention. The host cells can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic cell, such as a yeast cell, or a procaryotic cell such as a bacterial cell.

The present invention is further directed to isolated polypeptides and proteins encoded by ORFs of the present invention. A variety of methods, well known to those of skill in the art, routinely may be utilized to obtain any of the polypeptides and proteins of the present invention. For instance, polypeptides and proteins of the present invention having relatively short, simple amino acid sequences readily can be synthesized using commercially available automated peptide synthesizers. Polypeptides and proteins of the present invention also may be purified from bacterial cells which naturally produce the protein. Yet another alternative is to purify polypeptide and proteins of the present invention from cells which have been altered to express them.

The invention further provides methods of obtaining homologs of the fragments of the *Streptococcus pneumoniae* genome of the present invention and homologs of the proteins encoded by the ORFs of the present invention. Specifically, by using the nucleotide and amino acid sequences disclosed herein as



$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) = \frac{\partial L}{\partial x}$

a probe or as primers, and techniques such as PCR cloning and colony/plaque hybridization, one skilled in the art can obtain homologs.

The invention further provides antibodies which selectively bind polypeptides and proteins of the present invention. Such antibodies include both
5 monoclonal and polyclonal antibodies.

The invention further provides hybridomas which produce the above-described antibodies. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

The present invention further provides methods of identifying test samples
10 derived from cells which express one of the ORFs of the present invention, or a homolog thereof. Such methods comprise incubating a test sample with one or more of the antibodies of the present invention, or one or more of the DFs of the present invention, under conditions which allow a skilled artisan to determine if the sample contains the ORF or product produced therefrom.

15 In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the above-described assays.

Specifically, the invention provides a compartmentalized kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the antibodies, or one of the DFs of the present invention; and
20 (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of bound antibodies or hybridized DFs.

Using the isolated proteins of the present invention, the present invention further provides methods of obtaining and identifying agents capable of binding to
25 a polypeptide or protein encoded by one of the ORFs of the present invention. Specifically, such agents include, as further described below, antibodies, peptides, carbohydrates, pharmaceutical agents and the like. Such methods comprise steps of: (a) contacting an agent with an isolated protein encoded by one of the ORFs of the present invention; and (b) determining whether the agent binds to said protein.

30 The present genomic sequences of *Streptococcus pneumoniae* will be of great value to all laboratories working with this organism and for a variety of commercial purposes. Many fragments of the *Streptococcus pneumoniae* genome will be immediately identified by similarity searches against GenBank or protein databases and will be of immediate value to *Streptococcus pneumoniae* researchers



1. *Chlorophyll a* and *Chlorophyll b* were determined by the method of Arar and Collins (1971).

THIS PAGE BLANK (USPTO)

and for immediate commercial value for the production of proteins or to control gene expression.

The methodology and technology for elucidating extensive genomic sequences of bacterial and other genomes has and will greatly enhance the ability to analyze and understand chromosomal organization. In particular, sequenced contigs and genomes will provide the models for developing tools for the analysis of chromosome structure and function, including the ability to identify genes within large segments of genomic DNA, the structure, position, and spacing of regulatory elements, the identification of genes with potential industrial applications, and the ability to do comparative genomic and molecular phylogeny.

DESCRIPTION OF THE FIGURES

FIGURE 1 is a block diagram of a computer system (102) that can be used to implement computer-based systems of present invention.

FIGURE 2 is a schematic diagram depicting the data flow and computer programs used to collect, assemble, edit and annotate the contigs of the *Streptococcus pneumoniae* genome of the present invention. Both Macintosh and Unix platforms are used to handle the AB 373 and 377 sequence data files, largely as described in Kerlavage *et al.*, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, 585, IEEE Computer Society Press, Washington D.C. (1993). Factura (AB) is a Macintosh program designed for automatic vector sequence removal and end-trimming of sequence files. The program Loadis runs on a Macintosh platform and parses the feature data extracted from the sequence files by Factura to the Unix based *Streptococcus pneumoniae* relational database. Assembly of contigs (and whole genome sequences) is accomplished by retrieving a specific set of sequence files and their associated features using Extrseq, a Unix utility for retrieving sequences from an SQL database. The resulting sequence file is processed by seq_filter to trim portions of the sequences with more than 2% ambiguous nucleotides. The sequence files were assembled using TIGR Assembler, an assembly engine designed at The Institute for Genomic Research (TIGR) for rapid and accurate assembly of thousands of sequence fragments. The collection of contigs generated by the assembly step is loaded into the database with the lassie program. Identification of open reading

... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..

... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..
... ..

THIS PAGE BLANK (USPTO)

323

GTTTCGATG	GAGTTGTTGT	TGGAAATTGT	GTTTTTTCTA	CAACGTTAAA	GTTTTTCATCA	7620
CCGACAGCAC	AGACAAACTT	TGTACCGCCC	GCTTCCAAGC	TTCCATATAA	TTTTGTGATG	7680
ATAAACCTCT	TGTTTTTATT	TTCTTTATTA	TAGCATACTT	CGAAAGTCTA	AATGTCTCTA	7740
TTTTTTAGAT	TTTCTCTGT	AAATCTTACT	ATCTAATAAA	AACGAACAAA	CATGTCATTT	7800
GTTCCGTTTC	ACATTAGAGA	GGATTGATTA	GATTTTCACT	TCGATCACAG	CATCCCCCTT	7860
AGCAACTGAA	CCTGTTGCGA	CTGGAGCTAC	TGAAGCGTAG	TCACCTGTAT	TTGTAACGAT	7920
AACCATTGTT	GTATCATCAA	GTCCAGCTGC	AGCGATTTTG	TTTGAGTCAA	ATGTTCCAAG	7980
AACATCGCCA	GCTTTCACCT	TATTACCTTG	AGCAACTTTT	GTTTCAAAAC	CGTCACCGTT	8040
CATAGATACA	GTATCAATAC	CAACATGAAT	CAAAACTTCA	GCACCATTTC	TTGTTTTCAA	8100
ACCAAAAGCG	TGCCCTGTTG	GAAAGGCAAT	TGAAACTTCA	GCATCAGCTG	GTGCATAGAC	8160
CACGCCCTTG	CTTGGTTTCA	CAACGATACC	TTGTCCCATP	GCTCCACTTG	AGAAGACTGG	8220
GTCATTGACA	TCAGCAAGAG	CGACAACATC	ACCGACGATA	GGAGTTACAA	GTGTTTCATT	8280
TTGAAGAGCT	GCTGCCGCAA	CTTCTTCTTT	TTCTTCAGCC	ACTTCAGCTC	GTTTTGCACC	8340
TGCAGTTGCG	TCTACTTCAT	CTTCGTAACC	AAACATGTAA	GTAAGAGCAA	AACCAAGGGC	8400
AAATGATACA	GCTACCATAA	GAAGGTATTG	TGGAAGTTGT	CCGTTACCAA	CATAAAGCAT	8460
TGTACCAGGG	ATGATGGTGA	TACCATTACC	AGTACCAGCA	AGTCCAAGGA	TAGAAGCCAA	8520
TCCACCACCG	ATTGCACCAG	CAATCAATGA	AAGGAAGAAT	GGTTTACGGA	AGCGCAAGTT	8580
CACCCCGAAG	ATAGCAGGCT	CTGTAATACC	TAGGAAGGCA	GAAAGAGCAG	CCCGGAAAGC	8640
AAGTGTTC	AGTTTGGAT	TTTTTGTPTT	AACACCAACC	GCAACAGTAG	CAGCACCTTG	8700
AGCTGTCATA	GCAGCTGTGA	TGATAGCGTT	GAATGGGTTA	GCATGGTCAG	CAGCAAGTAA	8760
TTGCACTTCA	AGCAAGTTGA	AGATGTGGTG	CACACCTGAC	ACGACCATCA	ATTGGTGAAC	8820
CCCACCAATC	AAGAAACCAC	CAAGACCAAA	TGGCATGCTA	AGAATCGCTT	TTGTAGCAAT	8880
AAGGATGTAG	TTTTCAACAA	CGTGGAAGAC	TGGTCCAATG	ACAAAGAGTC	CAAGGATAGA	8940
CATGACCAAA	AGTGTACGGA	ATCGTGTAC	CAAGAGGTCA	ATGACATCTG	GAACAACCTG	9000
CCGACAGCTT	TTTCAAAATT	AGCTCCGACA	ACCCCGATGA	TGAAGGCTGG	AAGAACGGAA	9060
CCTTGCAAAC	CAACAACAGG	GATGAAACCA	AAGAAGTTCA	TGCTGTGTAC	TTCAACACCT	9120
TGAGCAACTG	CCCAAGCGTT	TGGAAGTGAG	CCAGAGACAA	GCATCATACC	AAGAACGATA	9180
CCAACGGCAG	GATTCCACC	AAATACACGG	AAGGTTGACC	ACACAACCAA	ACCTGGCAAG	9240
ATCATGAAGG	CTGTATCTGT	CAAGATTTGT	GTGTAAGTTG	CAAAGTCACC	TGGAAGTGCC	9300

THIS PAGE BLANK (USPTO)

What Is Claimed Is:

25 1. Computer readable medium having recorded thereon the nucleotide sequence depicted in SEQ ID NOS:1-391, a representative fragment thereof or a nucleotide sequence at least 95% identical to a nucleotide sequence depicted in SEQ ID NOS:1-391.

30 2. Computer readable medium having recorded thereon any one of the fragments of SEQ ID NOS:1-391 depicted in Tables 2 and 3 or a degenerate variant thereof.

35 3. The computer readable medium of claim 1, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

40 4. The computer readable medium of claim 3, wherein said medium is selected from the group consisting of a floppy disc, a hard disc, random access memory (RAM), read only memory (ROM), and CD-ROM.

 5. A computer-based system for identifying fragments of the *Streptococcus pneumoniae* genome of commercial importance comprising the following elements:

45 a) a data storage means comprising the nucleotide sequence of SEQ ID NOS:1-391, a representative fragment thereof, or a nucleotide sequence at least 95% identical to a nucleotide sequence of SEQ ID NOS:1-391;

 b) search means for comparing a target sequence to the nucleotide sequence of the data storage means of step (a) to identify homologous sequence(s), and

 c) retrieval means for obtaining said homologous sequence(s) of step (b).

50 6. A method for identifying commercially important nucleic acid fragments of the *Streptococcus pneumoniae* genome comprising the step of comparing a database comprising the nucleotide sequences depicted in SEQ ID NOS:1-391, a representative fragment thereof, or a nucleotide sequence at least 95% identical to a nucleotide sequence of SEQ ID NOS:1-391 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence is not randomly selected.

55

60 7. A method for identifying an expression modulating fragment of
Streptococcus pneumoniae genome comprising the step of comparing a database
comprising the nucleotide sequences depicted in SEQ ID NOS:1-391, a
representative fragment thereof, or a nucleotide sequence at least 95% identical to
the nucleotide sequence of SEQ ID NOS:1-391 with a target sequence to obtain a
nucleic acid molecule comprised of a complementary nucleotide sequence to said
65 target sequence, wherein said target sequence comprises sequences known to
regulate gene expression.

70 8. An isolated protein-encoding nucleic acid fragment of the *Streptococcus*
pneumoniae genome, wherein said fragment consists of the nucleotide sequence of
any one of the fragments of SEQ ID NOS:1-391 depicted in Tables 2 and 3, or a
degenerate variant thereof.

75 9. A vector comprising any one of the fragments of the *Streptococcus*
pneumoniae genome SEQ ID NOS:1-391 depicted in Tables 2 and 3 or a
degenerate variant thereof.

80 10. An isolated fragment of the *Streptococcus pneumoniae* genome,
wherein said fragment modulates the expression of an operably linked open reading
frame, wherein said fragment consists of the nucleotide sequence from about 10 to
200 bases in length which is 5' to any one of the open reading frames depicted in
Tables 2 and 3 or a degenerate variant thereof.

85 11. A vector comprising any one of the fragments of the *Streptococcus*
pneumoniae genome of claim 8.

12. An organism which has been altered to contain any one of the
fragments of the *Streptococcus pneumoniae* genome of claim 8.

90 13. An organism which has been altered to contain any one of the
fragments of the *Streptococcus pneumoniae* genome of claim 10.

14. A method for regulating the expression of a nucleic acid molecule comprising the step of covalently attaching to said nucleic acid molecule a nucleic acid molecule consisting of the nucleotide sequence from about 10 to 100 bases 5' to any one of the fragments of the *Streptococcus pneumoniae* genome depicted in SEQ ID NOS:1-391 and Tables 2 and 3 or a degenerate variant thereof.

15. An isolated nucleic acid molecule encoding a homolog of any of the fragments of the *Streptococcus pneumoniae* genome of SEQ ID NOS:1-391 and Tables 2 and 3, wherein said nucleic acid molecule is produced by a process comprising steps of:

a) screening a genomic DNA library using as a probe a target sequence defined by any of SEQ ID NOS:1-391 and Tables 2 and 3, including fragments thereof;

b) identifying members of said library which contain sequences that hybridize to said target sequence; and

c) isolating the nucleic acid molecules from said members identified in step (b).

16. An isolated DNA molecule encoding a homolog of any one of the fragments of the *Streptococcus pneumoniae* genome of SEQ ID NOS:1-391 and Tables 2 and 3, wherein said nucleic acid molecule is produced a process comprising steps of:

a) isolating mRNA, DNA, or cDNA produced from an organism;

b) amplifying nucleic acid molecules whose nucleotide sequence is homologous to amplification primers derived from said fragment of said *Streptococcus pneumoniae* genome to prime said amplification;

c) isolating said amplified sequences produced in step (b).

17. An isolated polypeptide encoded by any of the fragments of the *Streptococcus pneumoniae* genome of SEQ ID NOS:1-391 and depicted in Table 2 and 3 or by a degenerate variant of said fragments.

18. An isolated polynucleotide molecule encoding any one of the polypeptides of claim 17.

19. An antibody which selectively binds to any one of the polypeptides of claim 17.

20. A method for producing a polypeptide in a host cell comprising the steps of:

a) incubating a host containing a heterologous nucleic acid molecule whose nucleotide sequence consists of any one of the fragments of the *Streptococcus pneumoniae* genome of SEQ ID NOS:1-391 and depicted in Tables 2 and 3, under conditions where said heterologous nucleic acid molecule is expressed to produce said protein, and

b) isolating said protein.